REPORT DOCUMENTATION PAGE

Form Approved

OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
03-03-2015	Final	01-03-2012 – 28-02-2015

4. TITLE AND SUBTITLE

Situation Tracking in Large Data Streams

5a. CONTRACT NUMBER
FA2386-12-1-4054

5b. GRANT NUMBER
Grant AOARD-124054

5c. PROGRAM ELEMENT NUMBER
61102F

6. AUTHOR(S)

Andres Smith and Janet Wiles

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

ISSR, The University of Queensland
Queensland, Australia 4072

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

AOARD
UNIT 45002
APO AP 96338-5002

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/AFOSR/IOA(AOARD)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AOARD-124054

12. DISTRIBUTION/AVAILABILITY STATEMENT

Distribution A: Approved for public release. Distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Work on this project resulted in a system which: (a) identifies in real-time a schema definition of a scenario from an incomplete sample of a situation (the cue) plus a corpus of real-time unstructured and sparse time series data (such as social media or log records); (b) uses this schema to identify additional implicitly relevant data records to provide much greater recall of the scenario data, and quantifies the presence of schema elements in each data record; (c) identifies multiple different scenarios emerging from the data in response to a query; (d) quantifies the strength and stability of the time series of data records which contribute to each unfolding scenario; (e) generates a user interface which enhances situation awareness and minimizes the user's need for accurate prior knowledge, by enhancing the user's search terms and collating results; and (f) employs sound data science to construct data stories from statistically valid information, so non-analysts can make good decisions.

15. SUBJECT TERMS

Situation tracking, streaming data, Spcial media, Schema theory, Clustering, Word co-occurrence, Unstructured data, Real time analysis, Query, Story telling, Summarization

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Hiroshi Motoda, Ph. D.
U	U	U	SAR	11	19b. TELEPHONE NUMBER *(Include area code)* +81-42-511-2011

Situation Tracking in Large Data Streams

February, 2015

Principal Investigators:

Dr Andrew Smith
e-mail: a.smith7@uq.edu.au
Institution: The University of Queensland
Mailing Address: ISSR, The University of Queensland, Queensland, Australia, 4072
Phone: +61 400 337 247
FAX: +61 7 3346 7646
Prof Janet Wiles
e-mail: j.wiles@uq.edu.au
Institution: The University of Queensland
Mailing Address: The School of ITEE, The University of Queensland, Queensland, Australia, 4072
Phone: +61 7 336 52902

Period of Performance: March 1, 2012 – February 28, 2015

Abstract: Work on this project resulted in a system which: (a) identifies in real-time a schema definition of a scenario from an incomplete sample of a situation (the cue) plus a corpus of real-time unstructured and sparse time series data (such as social media or log records); (b) uses this schema to identify additional implicitly relevant data records to provide much greater recall of the scenario data, and quantifies the presence of schema elements in each data record; (c) identifies multiple different scenarios emerging from the data in response to a query; (d) quantifies the strength and stability of the time series of data records which contribute to each unfolding scenario; (e) generates a user interface which enhances situation awareness and minimizes the user's need for accurate prior knowledge, by enhancing the user's search terms and collating results; and (f) employs sound data science to construct data stories from statistically valid information, so non-analysts can make good decisions.

Introduction: The motivation for this research is that the information age has generated large quantities of certain sorts of weakly-structured time series data:

- Transactions: customer admin, sales, trading systems.
- Event logs: security monitors, systems monitors, web servers.
- Clinical monitoring: intensive care, long term therapy.
- Media: twitter, email, newswire.

Further, this sort of data poses real analytical challenges:

- A large set of possible data terms.
- The meaning of any term depends on the context at that time.
- Each record is sparse in that it only contains a small number of terms and a partial explanation of the situation, e.g. a tweet embedded in a Twitter thread.
- Simultaneous conversations can be interleaved - unreliable correlation between records at similar times.
- Noisy, e.g. spam.
- Non-stationary – changing statistical distributions of terms and term correlations.
- 'Stop words' (such as *the* and *is* in English) vary from source to source, time to time, language to language, and context to context.

One could summarise the problems with this sort of data in terms of the difficulty of acquiring an accurate situational awareness whenever one is attempting to interpret any given data record, in isolation. It is difficult to gain awareness of the relevant context for a specific datum, when there are so many contexts operating, and they change so quickly.

Most situational awareness technologies rely on forensic techniques and static models to identify emerging issues and trends; that is, "known" historical scenarios are used as the basis for modelling "unknown" emerging trends. The problem with forensic approaches is that they fail to capture new and emerging (previously undefined) scenarios: the "unknown unknowns". This research is exploring methods for ad hoc induction of a framework, or schema, to interpret a given data record.

Prior approaches to the problem of schema induction fall into three categories: (1) Computationally expensive techniques to extract general concept models, such as Latent Semantic Indexing. This technique generally extracts a single topic model, and does not extract a schema which is specifically relevant to each real-time query. It is generally expensive to rebuild the model to adapt to real-time data feeds. (2) Process Mining techniques to induce Petri Net models. This class of algorithms requires data records which must be pre-classified into scenario threads, and assumes stationary model statistics over the course of the time series. (3) Document clustering, which is computationally expensive as the number of records increases, and which does not generally allow a record to contribute to more than one cluster in a given model. Neither does document clustering work particularly well when used to cluster short fragments of text, such as tweets. Unless some other sematic model is applied, the feature vector for each document can only be constructed from the small number of terms present in the fragment. Such a sparse representation drawn from a much larger vocabulary severely restricts generalisation of the cluster model.

Analysis Method: Key features of the approach developed in this project include:

1. Extraction of an ad hoc concept schema in real time from small text units (e.g. sentences, tweets), using correlation analysis methods.
2. Allow multiple simultaneous state models to emerge. The need is to track multiple situations or events at the same time.
3. Creation of scenario charts based on mapping the time series of records matching the schema.
4. Early identification/prediction of emerging issues across multiple scenarios (stories) and data feeds.

Stage 1 of this work, completed in 2012-2013, was required to identify the meta-stable schemas of variables for situations present at that time in the data. The intention was to identify any emerging patterns much more deeply and accurately than simple term frequency and keyword search, because sparse records and highly contextual and correlated language render such approaches coarse and inaccurate. The system that emerged from the work during the first year followed the paradigm of Schema Theory [Bartlett, F.C. (1932). Remembering: A Study in Experimental and Social Psychology. Cambridge University Press]. Thus, a background schema was generated from the data in response to any user query, such as a word, a name, a time period, or a meta-data item.

Stage 2, completed in 2013-2014: It became clear from using the single schema system, developed in the first year, that this schema would often be composed of multiple scenarios within the one graph. This was particularly apparent when searching using a date range, as multiple situations are usually unfolding. So the first task for stage 2 was to find a fast method to decompose the background schema into clusters. The Newman-Girvan clusterer was found to be suitable [Newman, M. E. (2004). Analysis of weighted networks. Physical Review E, 70(5), 056131]. A major advantage of this algorithm is that it is non-parametric – it attempts to find the optimal partitioning of a network without the need to specify a threshold. This algorithm is also well suited to weighted networks. However, Newman-Girvan exclusively groups nodes, i.e. a node can only be in one cluster. This is not desirable for term networks, as it is often the case that a term is important evidence for multiple scenarios. Hence, we added a further algorithm to enlarge the resulting clusters to include terms non-exclusively.

This enlargement algorithm operates by finding for each node in the schema:

- A - the number of links from each node into any target cluster of nodes;

- B - the total number of nodes in the target cluster;
- C - the total number of links from each node.

Then, if A/sqrt(B x C) > 0.35, the node is added (non-exclusively) to the target cluster. The intent of this algorithm is to add nodes into clusters if they have a number of links into the node which are significant proportion of both the node's total link set, and the number of possible links into the cluster.

A further modification was found to be necessary in the last year of work. The search terms, because they conditionalise the whole schema, are too highly connected for satisfactory clustering. The issue is that major hubs in the schema, which are fairly uniformly connected to every other node, make it difficult for clustering algorithms to differentiate neighborhoods. As a result, leaving out the initial search terms from the schema passed to the Newman-Girvan clusterer produced a significant improvement in cluster discrimination.

These clusters were then used to extract weighted sub-graphs from the background schema. Each sub-graph acts as a second order classifier, and defines a scenario. The classifier is used to measure how relevant any text fragment is to the scenario. Hence we get a weighted thread or time-line of texts for each scenario. We also know which terms from the subgraph are activated by each text fragment, and this reveals the characteristic vocabulary of the scenario at any time. This will be useful for tracking changes in term-to-term relationships in a scenario.

We also explored measures of critical instability which could be applied to this time series (which is still quite sparse). The measure of fluctuation from Schiepek and Strunk was found to be reasonable [Schiepek, G., & Strunk, G. (2010). The identification of critical fluctuations and phase transitions in short term and coarse-grained time series—a method for the real-time monitoring of human change processes. *Biological cybernetics*, *102*(3), 197-207.] This measure uses the gradient of the time series between consecutive return points.

Various methods were tried to score the data records that matched each weighted sub-graph. It was found that the requirements differed between a metric used for ranking records by relevancy, and a metric used for time series quantification. The relevancy metric we employed sums the score for each pair of subgraph terms present in each record, multiplies by the number of pairs found to boost items which match more of the context, and then divides by the minimum of the number of terms in the data item or the number of terms in the subgraph, to reduce a bias towards long data items. This metric works well for thresholding and then ranking items for relevancy, but the value distribution tends towards being scale-free, with many very low values interspersed with steep spikes. For time series calculation, such as summing over time buckets and calculating the fluctuation score, it was found that a combination of Boolean scoring and relevancy scoring (as described above) produced a distribution with a tractable dynamic range. This was achieved by assigning a value of 0.5 to each data item found to be over threshold, and then adding the item's relevancy score, which is normalized to lie in the range of 0 to 0.5. The resulting score ranges between 0 and 1, and, when these scores are summed within time buckets, the time series fluctuations are much more tractable. More work is needed to characterize this combined distribution – it seems possible that the application of a suitable sigmoid weighting function would have a similar effect.

We judged the utility of this measure by working with text data sets with known points of crisis, such as the Iraq war news data and the Australian politics social media data.

Stage 3, completed in 2014-2015: Given the models and outputs generated in the earlier stages, the goal was then to select the best data presentations to support human cognitive strengths and weaknesses in order to enhance situation awareness.

The first human factor we address is the conundrum of how a user is to search for something in the data, when they don't know how it is optimally expressed in the data. Our situation tracking technology means that users do not have to struggle overly with guessing the appropriate search terms to find a situation which matches an information need. So long as their best guess is in the basin of the desired attractor in the data, the correlation engine will take them closer to the locally optimum result set.

The second human factor relates to the cognitive bias towards a compelling story, regardless of the quality of the data: "The confidence that individuals have in their beliefs depends mostly on the quality of the story they can tell about what they see, even if they see little." (Kahneman, D. 2011, pp87-88).
This leads us to believe that narrative is vital for understanding and acting, and that it is the job of the data scientist to not only account for data quality, base rates, and measurement of statistical correlation, but also to go further and present the conclusions as compelling stories. This will foster good decision making based on sound evidence. We synthesize statistically reliable narratives from the data and display these to the user - to maximize their engagement, understanding, and their ability to act on the information.

Applications: The system has been tested with multiple data corpora and live feeds:

1. A corpus of 100,000 sentences from 8 weeks of articles from The Australia newspaper containing the term Iraq over from12 Feb–28 April 2003 (before/after the invasion of Iraq).
2. A corpus of 35,000 tweets from July 2012 to October 2012 on the topic of the band *The Foo Fighters*.
3. Wikipedia snapshot (May 2014), containing around 200 million sentences.
4. A corpus of 846,000 tweets from 2011 on tech corporations.
5. The Enron email corpus, containing around 11 million sentences.
6. A live news feed of around 3 million sentences.
7. A feed of SEC filings.
8. Twitter live feeds: 10% and 1% feeds.

The system operates in this manner:

Query: the user enters a search query, which could be any data or meta-data fragment, including dates. Note that the cue can also be a data record itself. In that case, the system could identify the context activated by any encountered event record.
A (Schema): the system generates in real time a static term correlation matrix and from that can derive a concept tree – the graphical schema. Note that this schema contains any relevant terms or meta-data tags.
B (Clusters): our extended Newman-Girvan clusterer partitions the schema network into sub-graphs, one per scenario. Our intent here is to find collections of terms which travel together in order to describe a situation – these are our basic units of analysis, not individual words. This is essential given the correlated and non-linear nature of language usage. A single word is almost always ambiguous and rather devoid of meaning.
C (Story Cards): given the contextualised and narrative nature of language, we sought to communicate these aspects of a scenario to the user via a summary card for each scenario. These are created as follows: the text fragments that match each scenario are ranked by relevancy; the top ranked fragments which 'explain' the various term pairs in the sub-graph are assembled as summary points; we highlight the terms from the sub-graph in each summary point; the most frequently appearing terms from the sub-graph are displayed above the text examples.
D (Scenario Thread): This display, designed for providing access to the full story, uses the full list of terms from the sub-graph as filter control buttons (at top). Below, the matching text fragments are listed in either time order or relevancy order.
E (Time Series Export): Once all text fragments which match the scenario have been scored, the scores are bucketed into time intervals, along with the frequency of occurrence of active sub-graph terms in each interval. The Fluctuation measure of instability is also calculated.

An Example:

Consider the query: "Hornet" on Corpus 1 (the Iraq war newspaper articles).

Here is the sequence of results:

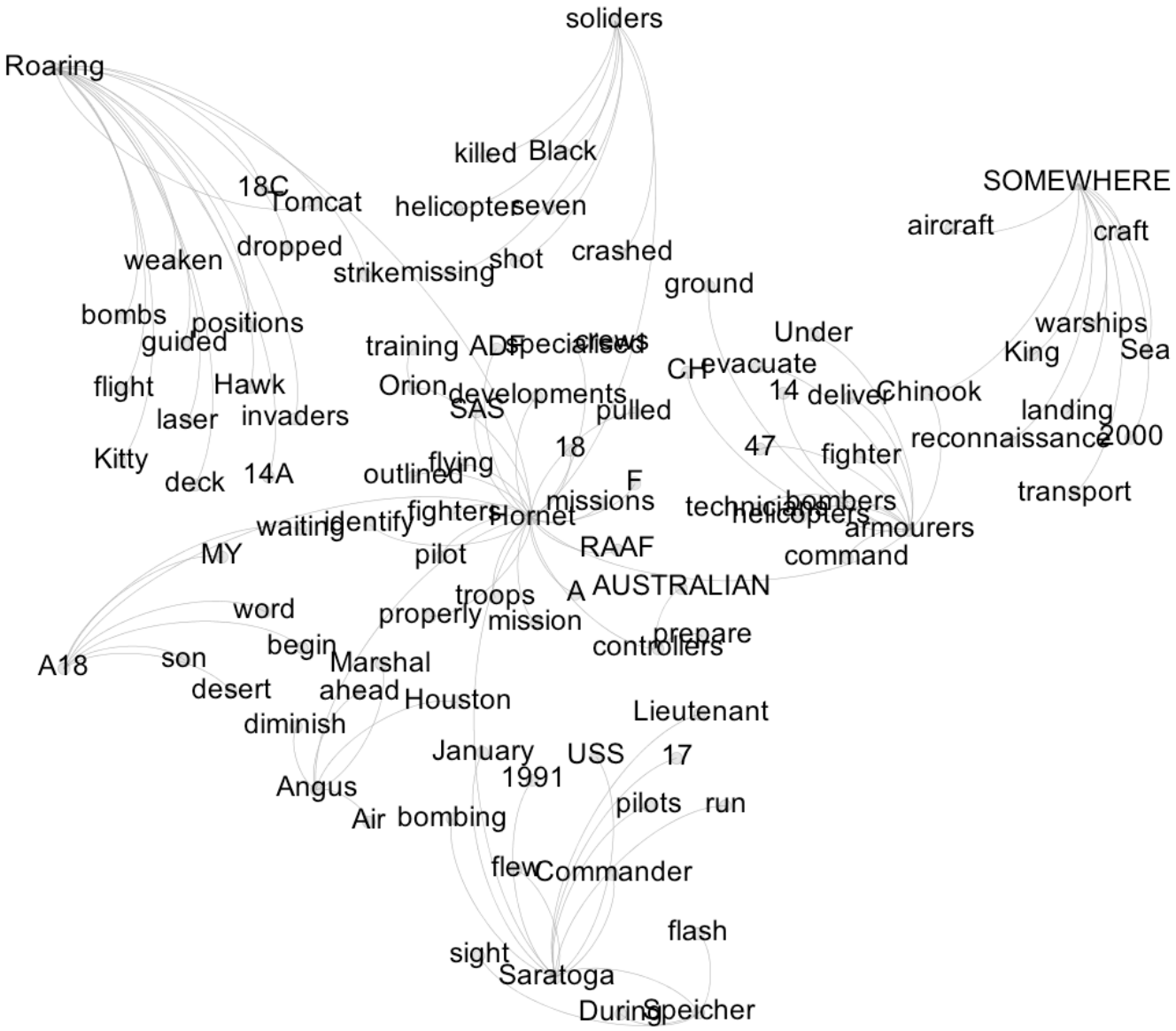

Figure 1: Background schema found from the query "Hornet" (with links simplified using a spanning tree)

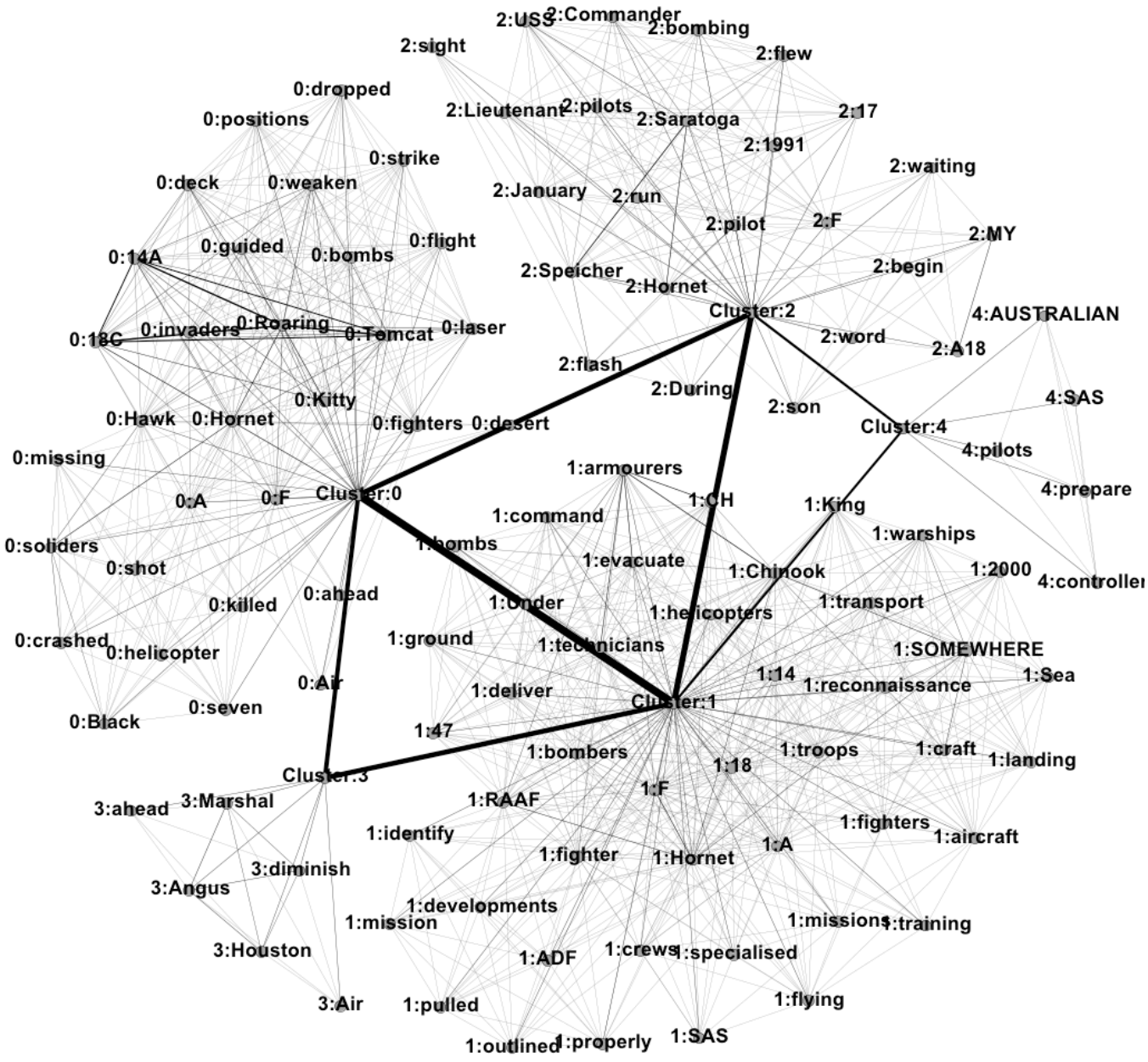

Figure 2: Schema clustered into distinct scenarios using non-exclusive Newman-Girvan (all links shown within each cluster) from the query "Hornet"

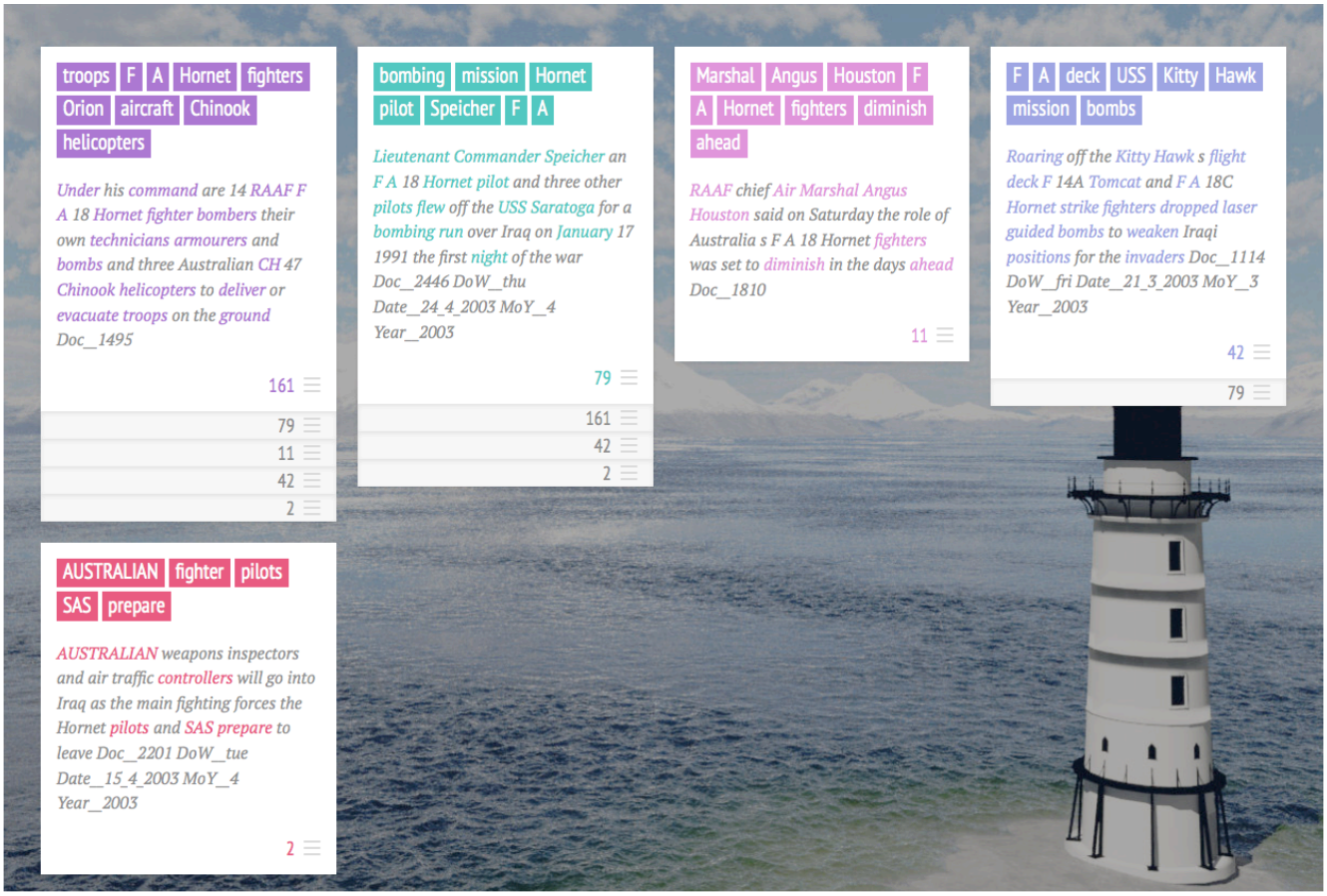

Figure 3: Summary cards for scenarios (clusters) found from the query "Hornet". Please note that the UI/tool used to visualise this data is owned and copyright by Hypermancer SIA, 2014

Summary Cards: We generate summary cards (see Figure 3) that are designed to quickly communicate the meaning of each scenario to a user. The goal is to present a data narrative, which is grounded in representative statistics but nevertheless communicates the central story to the user in a compelling manner. The "headlines" of keywords display those key terms from the cluster that appear in the top ranked matching text records. The text shown below the headline on the card is the top ranked text record matching the scenario, with the schema terms highlighted.
On the "flip-side" of each card, the user can inspect full list of terms in the cluster, and all the matching text records above a selected relevancy threshold (see Figure 4).

Figure 4: Text records that match a "pilot" scenario. Please note that the UI/tool used to visualise this data is owned and copyright by Hypermancer SIA, 2014

The weights of the text records which match each scenario can also produce a time series, and the stability of this can be estimated using the fluctuation metric derived from (Schiepek, G., & Strunk, G. 2010). See Figure 5.

Date	Cluster	Top Terms	Strength	Instability
Wed Feb 12 10:00:00 2003	2		0	
Thu Feb 13 10:00:00 2003	2		0	0
Fri Feb 14 10:00:00 2003	2	bombing pilots 1991 During	0.220260958	0.22026
Mon Feb 17 10:00:00 2003	2		0	0.22026
Tue Feb 18 10:00:00 2003	2		0	0
Wed Feb 19 10:00:00 2003	2		0	0
Thu Feb 20 10:00:00 2003	2		0	0
Fri Feb 21 10:00:00 2003	2		0	0
Mon Feb 24 10:00:00 2003	2		0	0
Tue Feb 25 10:00:00 2003	2	flew 1991	0.220468315	0.22046
Wed Feb 26 10:00:00 2003	2		0	0.22046
Thu Feb 27 10:00:00 2003	2		0	0
Fri Feb 28 10:00:00 2003	2	1991 positions bombing	0.22014214	0.22014
Mon Mar 3 10:00:00 2003	2		0	0.22014
Tue Mar 4 10:00:00 2003	2	F word	0.110257843	
Wed Mar 5 10:00:00 2003	2	Commander Lieutenant 1991 January During	0.330943664	0.16547
Thu Mar 6 10:00:00 2003	2		0	0.33094
Fri Mar 7 10:00:00 2003	2	Kitty Hawk aircraft USS January guided	0.11018386	0.11018
Mon Mar 10 10:00:00 2003	2		0	0.11018
Tue Mar 11 10:00:00 2003	2		0	0
Wed Mar 12 10:00:00 2003	2	Commander Lieutenant 1991 bombing During	0.220503581	0.2205
Thu Mar 13 10:00:00 2003	2		0	0.2205
Fri Mar 14 10:00:00 2003	2	1991 Lieutenant Air training F A 18	0.330299767	0.33029
Mon Mar 17 10:00:00 2003	2		0	0.33029
Tue Mar 18 10:00:00 2003	2	F pilot RAAF word son MY Hornet waiting begin desert A18 18	1	1
Wed Mar 19 10:00:00 2003	2		0	1
Thu Mar 20 10:00:00 2003	2	pilots F strike USS Tomcat pilot aircraft Kitty ahead 14A missions Commander	0.551562814	0.55156
Fri Mar 21 10:00:00 2003	2	F run pilot Lieutenant 14 ground bombing troops mission USS bombs deck	0.221066071	0.33049
Mon Mar 24 10:00:00 2003	2	F mission 18 pilot A Hornet developments properly fighter RAAF ADF identify	0.332225666	0.11115
Tue Mar 25 10:00:00 2003	2		0	0.33222
Wed Mar 26 10:00:00 2003	2	Lieutenant F A 18 pilots bombing aircraft fighter pilot SAS flew RAAF	0.99209933	0.99209
Thu Mar 27 10:00:00 2003	2	killed USS 2000 17	0.220332658	0.77176
Fri Mar 28 10:00:00 2003	2	Lieutenant Commander aircraft USS Kitty Hawk fighters command bombing pilots	0.551694112	0.33136
Sat Mar 29 10:00:00 2003	2		0	0.55169
Mon Mar 31 10:00:00 2003	2		0	0
Tue Apr 1 10:00:00 2003	2		0	0
Wed Apr 2 10:00:00 2003	2	ground run Lieutenant	0.110101589	0.1101
Thu Apr 3 10:00:00 2003	2		0	0.1101
Fri Apr 4 10:00:00 2003	2	F A 18 helicopter soliders missing Hornet seven Hawk Black pilot shot	0.662014545	0.66201
Mon Apr 7 10:00:00 2003	2	bombing F fighter ahead flew pilots A ground 18	0.330652914	
Tue Apr 8 10:00:00 2003	2		0	0.331
Wed Apr 9 10:00:00 2003	2	bombing 1991 command mission 17 bombs	0.220127754	0.22012
Thu Apr 10 10:00:00 2003	2		0	0.22012
Fri Apr 11 10:00:00 2003	2	flew 1991	0.110234157	
Mon Apr 14 10:00:00 2003	2	F A 18 bombing pilots	0.11029201	
Tue Apr 15 10:00:00 2003	2	controllers AUSTRALIAN pilots SAS Hornet prepare	0.110537924	0.03684
Wed Apr 16 10:00:00 2003	2		0	0.11053
Fri Apr 18 10:00:00 2003	2	Lieutenant positions SAS 17	0.110081275	
Mon Apr 21 10:00:00 2003	2	waiting pilot	0.110326717	0.05516
Tue Apr 22 10:00:00 2003	2	F 17	0.110151487	
Wed Apr 23 10:00:00 2003	2		0	0.05516
Thu Apr 24 10:00:00 2003	2	Speicher pilot Hornet Lieutenant Commander F USS 17 flew run 18 1991	0.581298948	0.58129
Mon Apr 28 10:00:00 2003	2		0	0.29064

Figure 5: Time series of text record cluster terms, relevancy scores, and instability scores for a selected scenario – in this case the "pilot" scenario found from the query "Hornet". For reference, the invasion of Iraq lasted from 19 March 2003 to 1 May 2003.

List of Publications and Significant Collaborations:

A schema generation process and system. Australian Provisional Patent Application No. 2013900363. Filing date: 5 February, 2013. Smith, A. E.
A schema generation process and system. International Patent Application No. PCT/AU2014/000081. Filing date: 5 February, 2014. Smith, A. E.

Bibliography

Bartlett, F.C. 1932. *Remembering: A Study in Experimental and Social Psychology*. Cambridge University Press.

Coulson, S. 2001. *Semantic leaps: frame-shifting and conceptual blending in meaning construction*. Cambridge University Press.

Humphreys, M. S, Bain, J. D, & Pike, R. 1989: Different Ways to Cue a Coherent Memory System: A Theory for Episodic, Semantic, and Procedural Tasks. *Psychological Review,* 96(2), 208-233

Kahneman, D. 2011. *Thinking, fast and slow*. Macmillan.

Lakoff, G. 1987. *Women, Fire, and Dangerous Things*. The University of Chicago Press.

Langacker, R.W. 1987. *Foundations of Cognitive Grammar: Theoretical prerequisites*. Stanford University Press.

Newman, M. E. 2004. *Analysis of weighted networks*. Physical Review E, 70(5), 056131.

Schiepek, G., & Strunk, G. 2010. *The identification of critical fluctuations and phase transitions in short term and coarse-grained time series—a method for the real-time monitoring of human change processes*. Biological cybernetics, 102(3), 197-207.

Tulving, E. 1985. How Many Memory Systems Are There? *American Psychologist*, Vol. 40, No. 4, 385-398.

Wiles, J, Humphreys, M.S., Bain, J.D., and Dennis, S. 1990. Control Processes and Cue Combinations in a connectionist model of human memory. Department of Computer Science Technical Report No. 186, University of Queensland. (unpublished).

Wittgenstein, L. 1953 (trans. 1958). *Philosophical Investigations*. Basil Blackwell, 1958.